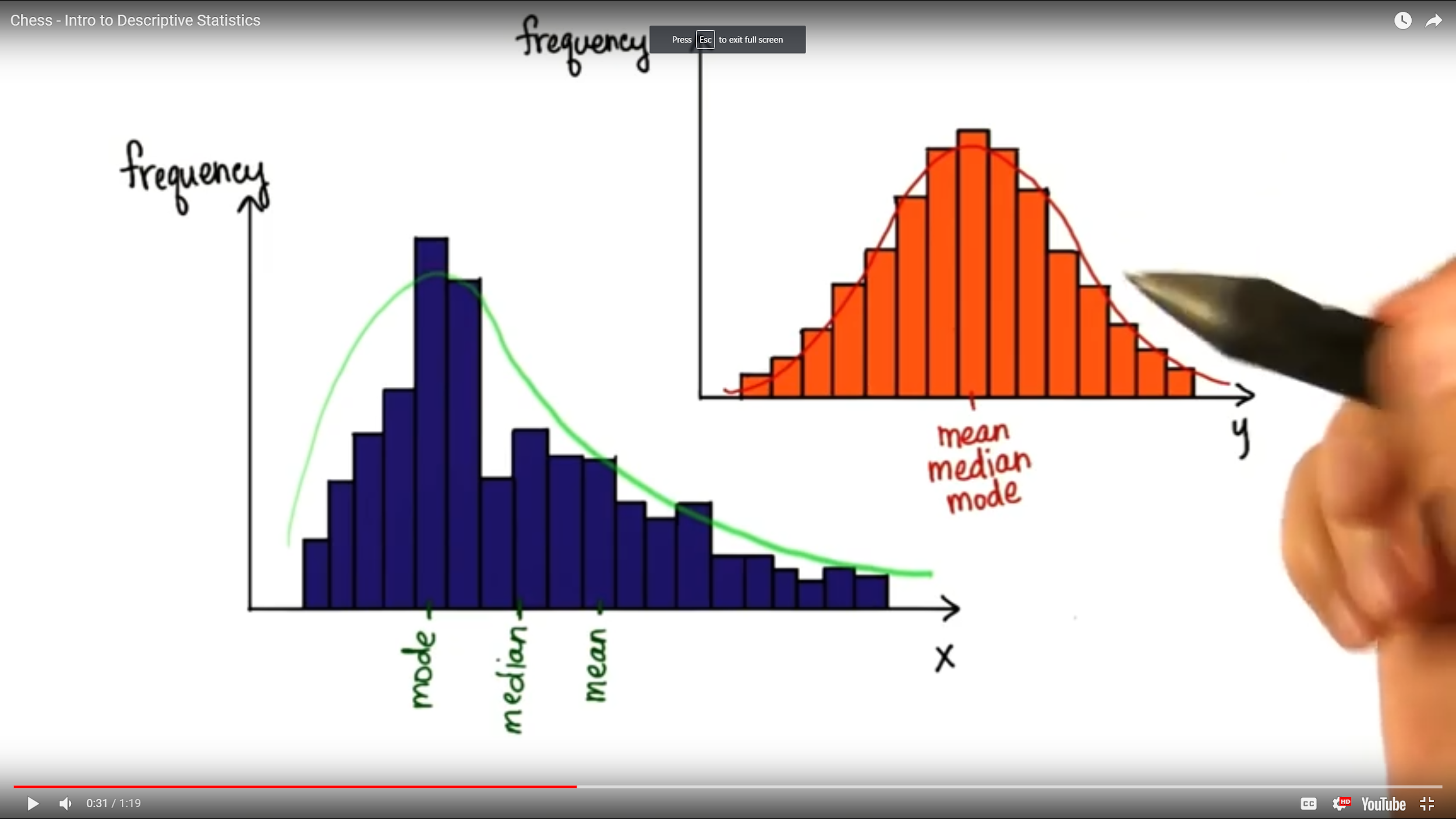
BS – Important Concepts Standardizing – Lessons 15, 16, 17

1. Visualize data set with histograms in order to analyze shape:

→ to think critically about the mean, median and the mode to describe the data set;



→ in a skewed distribution, the mean, median and the mode differ from each other and the median might be more useful than the mean;

→ in a normal distribution, the mean, median and the mode are approximately equal;

1. The proportion of data values, less than or greater than a certain value in the data set

is important:

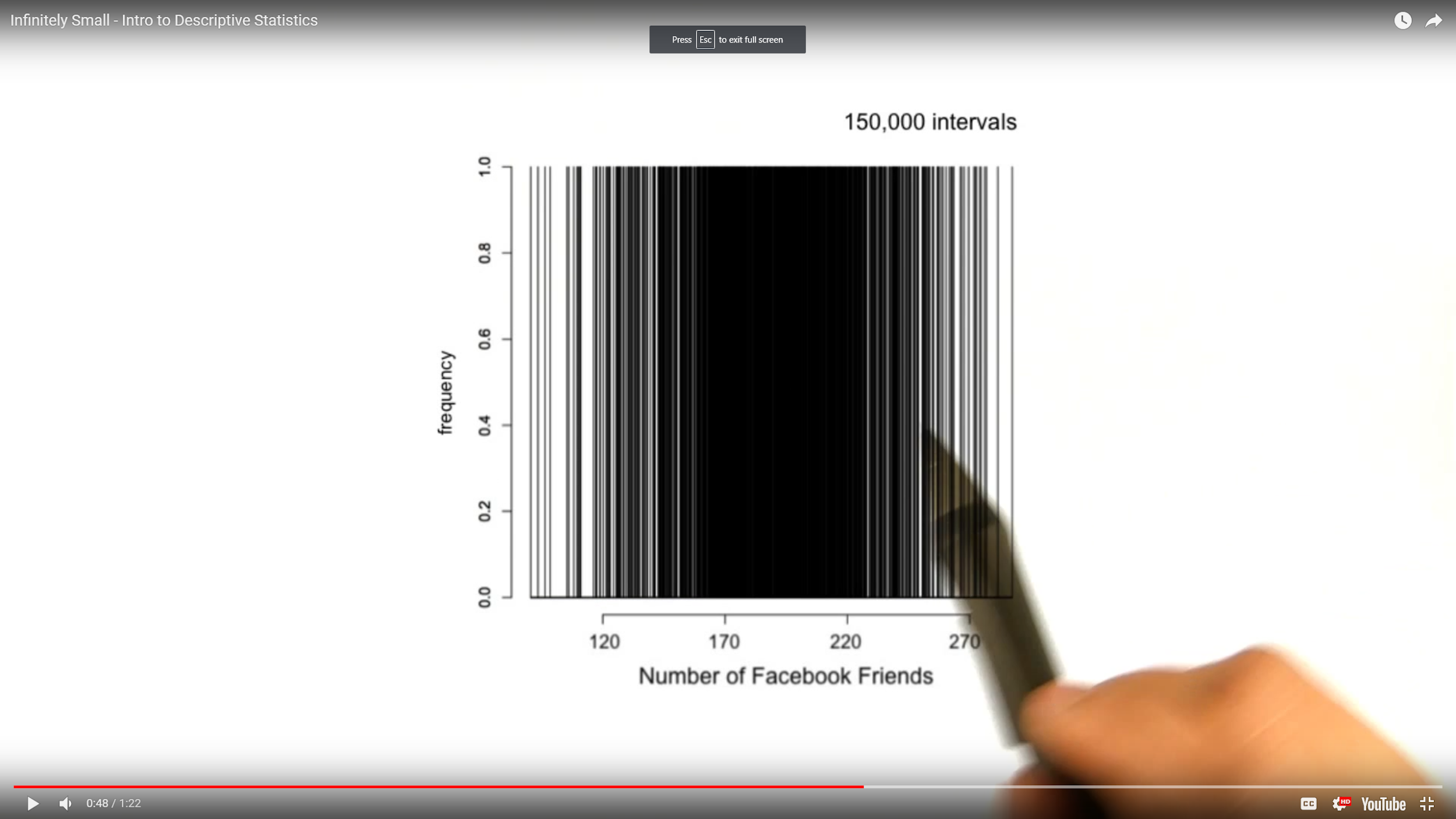
→ use relative frequencies and convert all absolute frequencies to a proportion;

→ the relative frequency distribution looks almost exactly the same as the absolute frequency distribution;

→ we want to determine the proportion of values less than or greater than certain numbers to compare scores in the distribution;

→ a smaller bin size will allow more detail: Ideally, we would want a smaller bin size as possible, infinitely small → but then the frequency on y-axis is getting smaller → in this case,

frequency = 0 or 1 => we lose the shape of the distribution



1. Continuous Distribution:

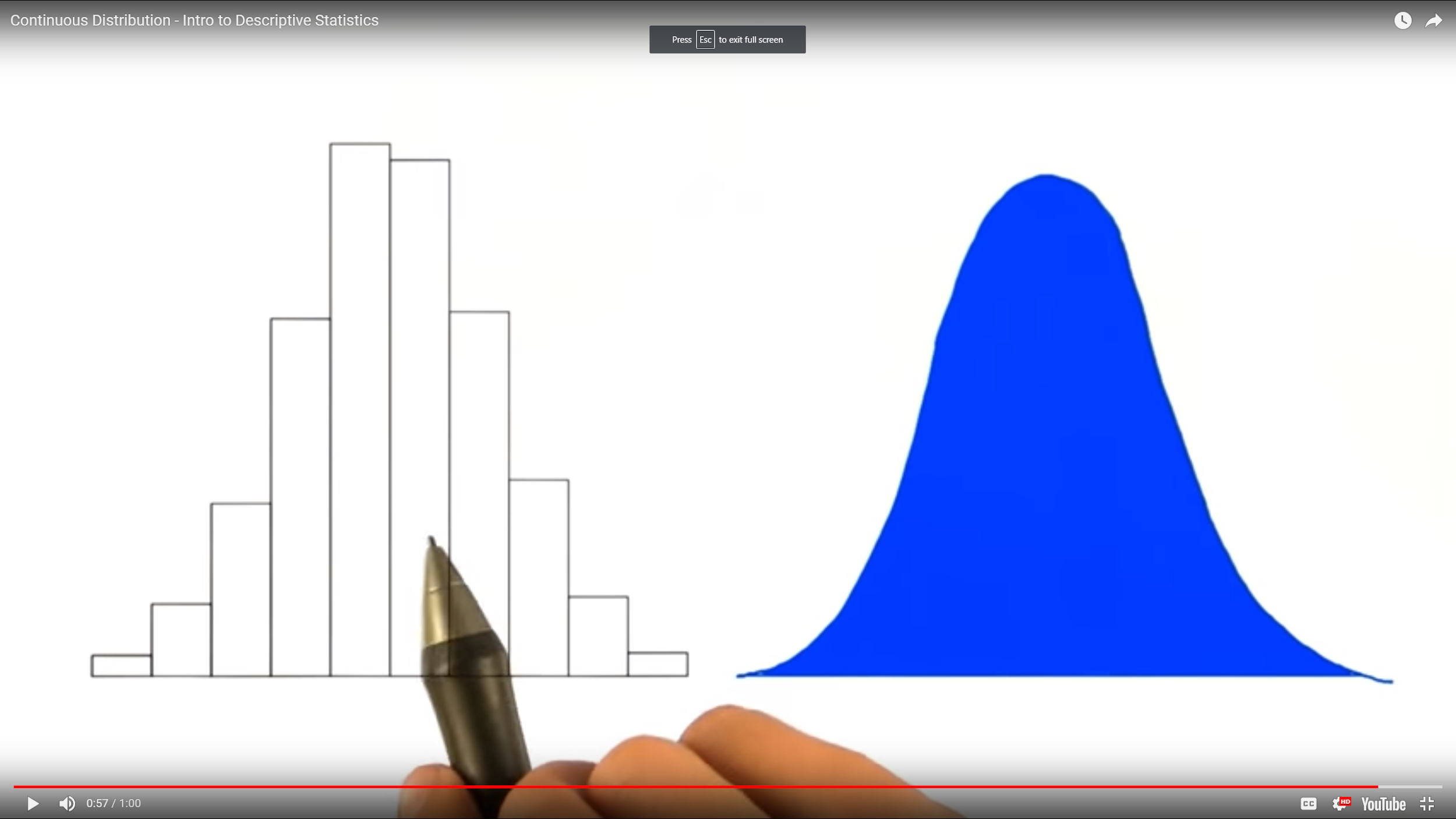
→ we want to locate exactly data values, relative to the rest of the distribution => the answer is using a theoretical model for our distributions;

→ smooth curve that uses relative frequencies;

→ it can be described with an equation;

→ allows us to calculate the proportion between any two values on the x axis;

→ the area under the curve is 1 or 100%;



1. Theoretical Normal distributions:

→ stretched, skinny;

→ mean = median = mode;

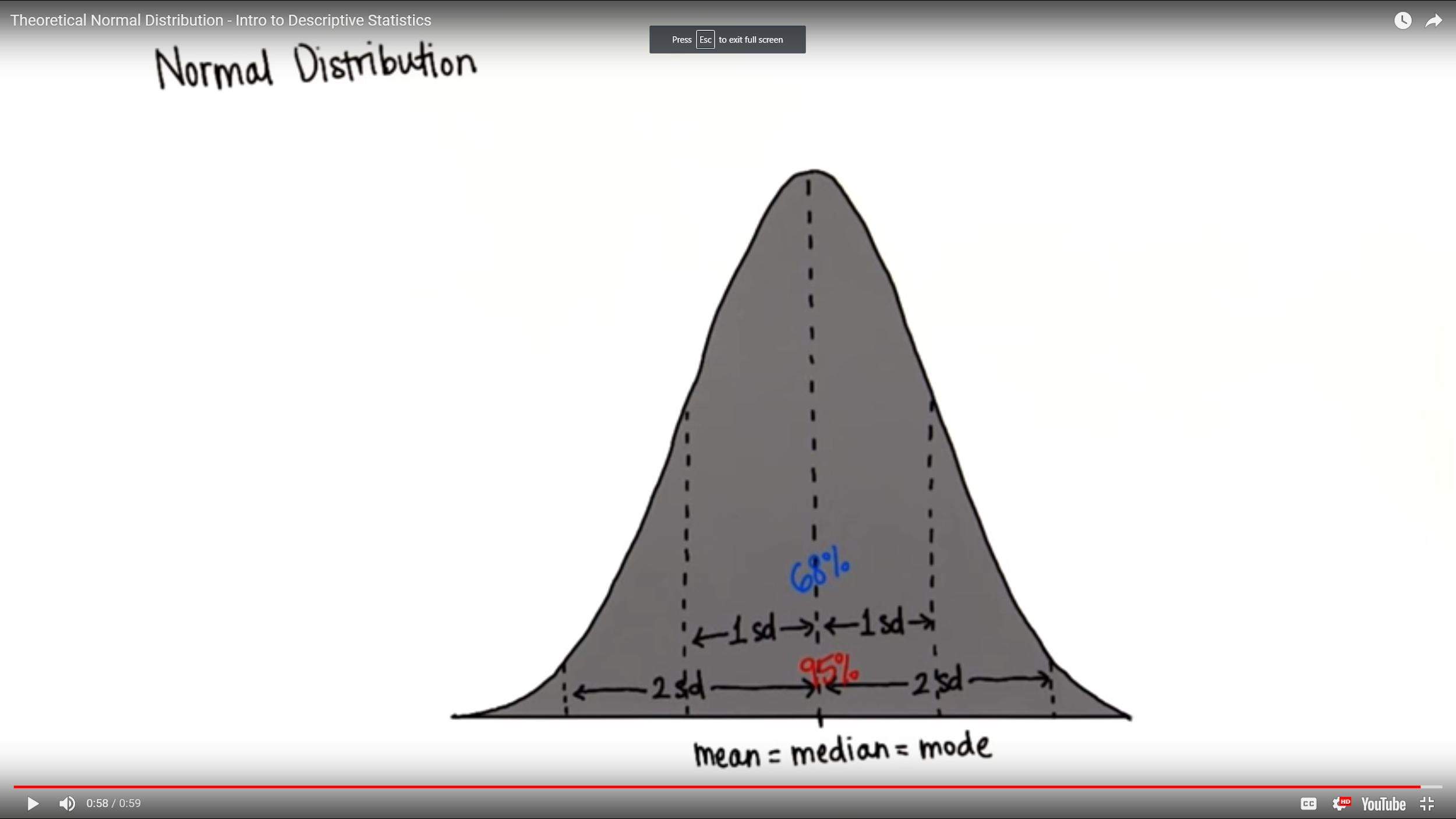
→ perfectly symmetrical;

→ they approximate our real distributions;

→ most of the data is in the middle;

→ 68% off all data falls within 1 standard deviation of the mean;

→ 95% off all data falls within 1 standard deviations of the mean

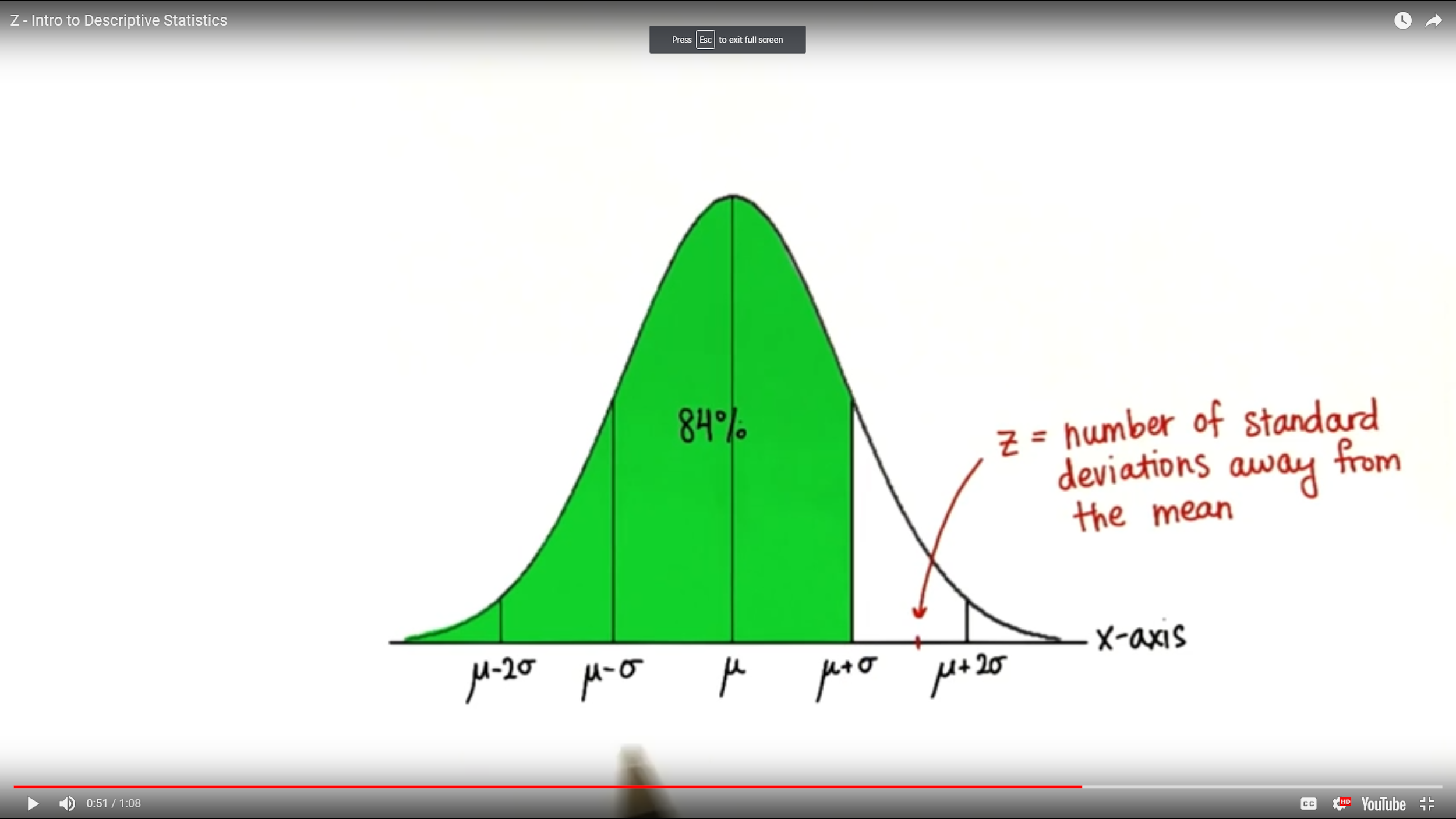


1. Z-Score: Given an observed value x, the Z score finds the number of Standard

deviations x is away from the mean.

→ the location of particular values on the x-axis described in terms of standard deviations;

→ whatever the score, it can be converted to a value = to the number of standard deviations away from the mean = Z;

→ by converting score to Z we can know the percent less than or greater than that value;

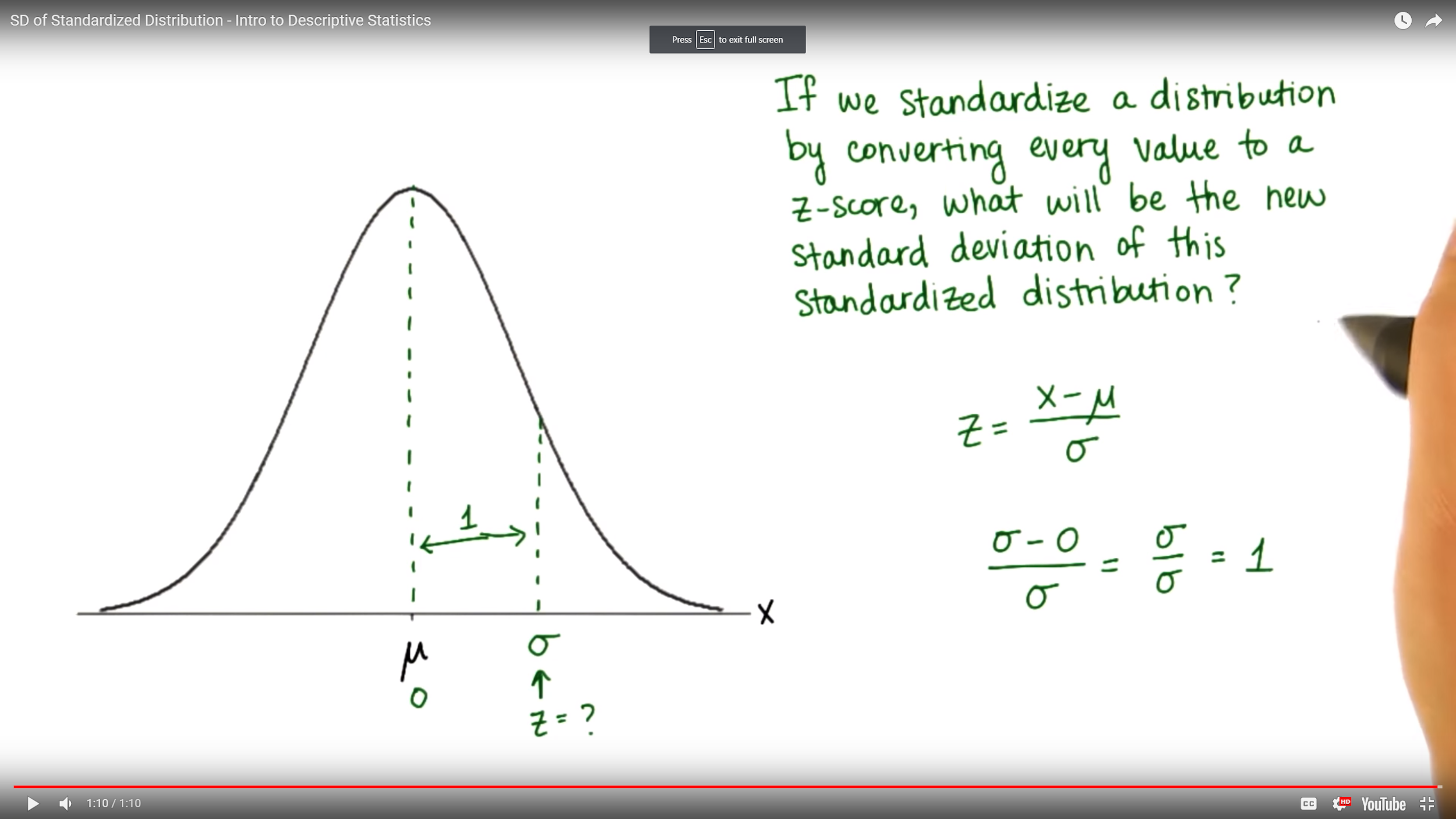
→ if the score is below/above the mean;

→ standardize any score on x-axis => standardize the distribution

→ the new mean of the standardized distribution is 0 because we shift the standardized distribution by subtracting the mean, centering at 0;

→ the z-score of the mean is 0;

→ the new standard deviation of the standardized distribution is 1 because the z score of the Standard deviation (sigma) is 1 => any value that is 1 standard deviation away from the mean will be one after we standardize it



1. Standardizing distributions:

→ compare distributions by looking at them on the same scale;

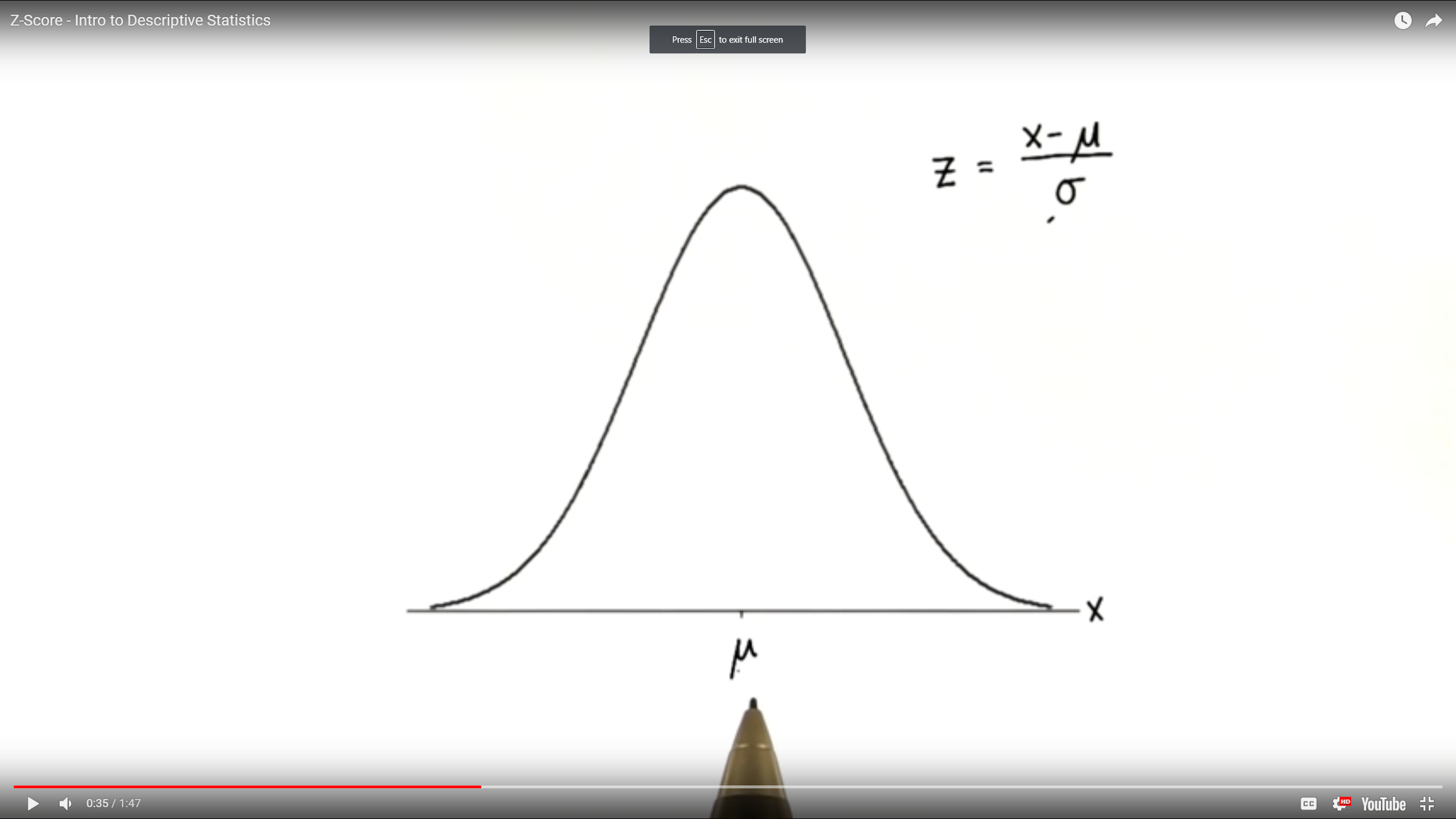
→ in terms of their unique standard deviations;

→ 0 is the reference point;

→ standardized scores show us the proportion that have a lower or higher score in that distribution;

1. Formula to find the number of Standard deviations each value is away from the

mean - Z-score:



1. A negative Z-score:

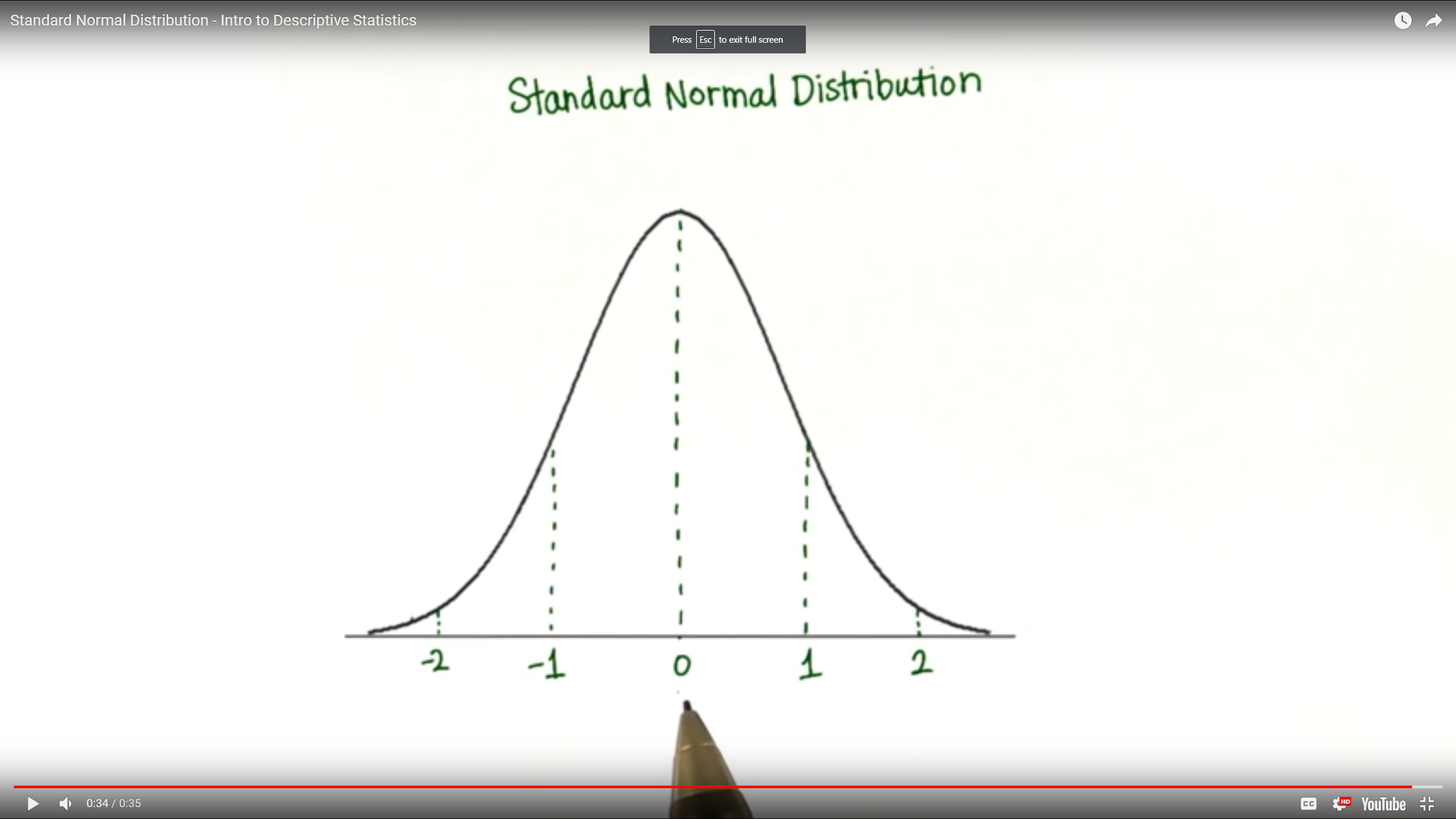
→ the original value is less than the mean;

→ the original value minus the mean is negative;

1. Standard Normal Distribution:

→ when we have any normal distribution we can standardize it by first subtracting the mean (shifting it to 0) and then dividing by the standard deviation which makes the Standard Deviation = 1;

→ every value in the data set is written in terms of how many standard deviations it is away from the mean



1. Example: Convert the number of FB friends for x with known mean and sigma to a

Popularity Score with known mean and sigma:

→ calculate the z-score for x (convert to z-score)

→ convert it to new popularity score = z-score for x \* sigma of Popularity Score + mean of Popularity Score

